Эконометрика, 2020-2021, 3 модуль
Семинары 5-6
8, 15.02.21
Для Группы Э_Б2018_Э_3
Семинарист О.А.Демидова
Модели бинарного выбора

Демидова, Малахов, 11.5

Задание 11.5.

База данных rlms14 (файл rlms14.dta)

В файле rlms14.dta приведены данные о заработной плате, возрасте, образовании, профессиональной принадлежности россиян из базы данных RLMS, 14 раунд опроса, проводившийся в 2005 г.

Переменные:

income – ответ на вопрос: «Сколько денег в течение последних 30 дней Вы получили по основному месту работы после вычета налогов и отчислений?»

wage - ответ на вопрос: «Ваша среднемесячная зарплата за последние 12 месяцев?»

psu – первичная единица отбора (1 – Санкт – Петербург, …, 38 – Амурская обл.),

marst – семейное положение (1 – никогда в браке не состояли, 2- состоите в зарегистрированном браке, 3 – живете вместе, но не зарегистрированы, 4 – разведены и в браке не состоитн, 5 – вдовец/вдова),

occup – профессиональная группа (1- военнослужащие, 2 – законодатели, крупые чиновники, управленцы, 3 – профессионалы с высшим образованием, 4 – конторские служащие, 5 – занятые в сфере обслуживания, 6 – квалифицированные с/х работники, 7 – ремесленники, 8 – промышленные рабочие),

educ – образование (1 – 0 классов школы, 23 – аспирантура и т.п. с дипломом),

diplom – законченное образование (1 – 0-6 классов, 1-2 – незаконченное среднее образование, 3 - законченное среднее образование, 4 - законченное среднее образование, 5 - законченное среднее специальное образование, 6 - законченное высшее образование и выше),

gender – пол респондента (1 – мужской, 2 – женский),

birth_year – год рождения,

industry – ответ на вопрос: «в какой отрасли Вы работаете?» (1 – легкая и пищевая промышленность, …),

beginning – ответ на вопрос: «с какого года Вы работаете на этом предприятии?»

boss – ответ на вопрос: «У Вас есть подчиненные на этой работе?» (1 –да, 2 – нет),

subordinates – ответ на вопрос: «Сколько у Вас подчиненных?»

duration_weekh - ответ на вопрос: «Сколько часов в среднем продолжается Ваша обычная рабочая неделя?»

life_sat - ответ на вопрос: «Насколько Вы удовлетворены или не удовлетворены Вашей жизнью в целом?» (1 - «Полностью удовлетворены», …, 5 – «Совсем не удовлетворены»),

job_sat - ответ на вопрос: «Насколько Вы удовлетворены или не удовлетворены Вашей работой в целом?» (1 - «Полностью удовлетворены», 2 - «Скорее удовлетворены», 3 – «И да, и нет», 4 – «Не очень удовлетворены», 5 – «Совсем не удовлетворены»),

job_career - ответ на вопрос: «Насколько Вы удовлетворены или не удовлетворены возможностями Вашего профессионального роста?» (1 - «Полностью удовлетворены», …,5 – «Совсем не удовлетворены»),

children - ответ на вопрос: «Сколько у Вас детей?»,

child_less18 - ответ на вопрос: «Сколько у Вас детей моложе 18 лет?»,

По данным файла rlms14.dta с помощью статистического пакета STATA попытайтесь выявить факторы, влияющие на удовлетворенность россиян жизнью в целом, работой, возможностями профессионального роста.

Например, для выявления факторов, влияющих на удовлетворенность россиян жизнью в целом

1) Создайте зависимую переменную life, равную 1 для тех, кто на вопрос «Насколько Вы удовлетворены или не удовлетворены Вашей жизнью в целом?» ответили «Полностью удовлетворены» или «Скорее удовлетворены» (т.е. для кого значения переменной life_sat равно 1 или 2).
2) Оцените логит и пробит модели с зависимой переменной life, объясняющие переменные выберите самостоятельно (при необходимости создайте новые, например

   возраст age = 2005 – birth_year и т.п.), Вычислите предельные эффекты объясняющих факторов в выбранной Вами точке. Дайте интерпретацию полученным результатам.

   Используйте статистический пакет STATA.

Методические указания.

Полезные команды:

```
set more off
gen life = 0
replace life =1 if job_sat <3
gen age = 2005- birth_year
tab educ
tab  diplom
logit life gender income children age i.diplom
label list psu
logit life gender income children age i.diplom if psu==2
gen higheduc=0
replace higheduc=1 if diplom==6
```

probit life gender income children age higheduc if psu==2

mfx

sum gender income children age higheduc

mfx, at (1 5000 0 25 1)

probit life income children age higheduc if psu==2 & gender ==2

Дополнительные задания.

3) Оцените качество оценки моделей

Полезные команды:

estat classification

estat gof

estat classification, cutoff(0.4)

4) Постройте график кривых чувствительности и специфичности

Полезная команда:

lsens

5) Постройте график ROC – кривой и вычислите площадь под ней

Полезная команда:

lroc

Демидова, Малахов, задачи 11.1-11.4, задание 11.3

Задача 11.1.

Исследователя интересует зависимость вероятности найти работу от уровня образования индивидуума. Введя в качестве зависимой переменную EMP, равную 1 для работающих и 0 для неработающих и S – количество лет обучения в качестве объясняющей, он оценил логит – модель:

$$P\{EMP_i = 1\} = \frac{1}{1+\exp\{-Z_i\}}, \quad Z_i = \underset{(0.242)}{-1.006} + \underset{(0.018)}{0.148}\,S,$$

Оцените предельный эффект объясняющего фактора для среднего значения переменной $S = 13.5$.

Решение:

$$Z(\bar{s}) = -1.006 + 0.148 \cdot 13.5 = 0.992$$

$$\exp(-z(\bar{s})) = \exp(-0{,}992) = 0{,}371$$

$$f(z) = \frac{e^{-z}}{(1+e^{-z})^2}$$

$$f(z(\bar{s})) = \frac{0{,}371}{(1+0{,}371)^2} = 0{,}1973$$

$$\frac{\partial p(\bar{s})}{\partial s} = \frac{\partial p}{\partial z} \cdot \frac{\partial z}{\partial s} = f(z(\bar{s})) \cdot \hat{\beta}_j = 0{,}148 \cdot 0{,}1973 = 0{,}029$$

Интерпретация: при увеличении длительности обучения (S) на 1 год для индивида со «средними» способностями вероятность найти работу увеличивается на 2.9%.

Задача 11.2.

По наблюдениям для 570 индивидуумов оценена зависимость получения школьником аттестата от обобщенной оценки результатов тестов X. Переменная Y принимает значение 1, если аттестат был получен и 0 в противном случае.

Оцененные модели имеют следующий вид:

Логит: $P\{Y_i = 1\} = \dfrac{1}{1 + \exp\{-Z_i\}}$, $Z_i = \underset{(0.865)}{-5.004} + \underset{(0.021)}{0.1666}\, X$,

Пробит: $P\{Y_i = 1\} = F(Z_i)$, $F(z) = \dfrac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-t^2/2} dt$, $Z = \underset{(0.083)}{-2.7} + \underset{(0.0117)}{0.53}\, X$

Дайте экономическую интерпретацию полученным результатам для логит и пробит моделей. Найдите предельный эффект объясняющего фактора в точке $\overline{X} = 50.15$.

Решение:

Найдем предельный эффект в логит модели:

$$z(\bar{x}) = -5{,}004 + 0{,}1666 \cdot 50{,}15 = 3{,}351$$

$$\exp(-z(\bar{s})) = \exp(-0{,}992) = 0{,}0351$$

$$f(z) = \frac{e^{-z}}{(1 + e^{-z})^2}$$

$$f(z(\bar{x})) = \frac{0{,}0351}{(1 + 0{,}0351)^2} = 0{,}033$$

$$\frac{\partial p(\bar{x})}{\partial x} = \frac{\partial p}{\partial z} \cdot \frac{\partial z}{\partial x} = f(z(\bar{x})) \cdot \hat{\beta}_j = 0{,}033 \cdot 0{,}1666 \approx 0{,}0055$$

При увеличении обобщенной оценки результатов тестов на 1 балл для индивида с обобщенной оценкой равной 50,15 вероятность получения аттестата увеличивается на 0,55%.

Найдем предельный эффект в пробит модели:

$$z(\bar{x}) = -2{,}7 + 0{,}53 \cdot 50{,}15 = 23{,}88$$

$$f(z(\bar{x})) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \approx 0$$

$$\frac{\partial p(\bar{x})}{\partial x} = \frac{\partial p}{\partial z} \cdot \frac{\partial z}{\partial x} = f(z(\bar{x})) \cdot \hat{\beta}_j \approx 0 \cdot 0{,}53 \approx 0$$

При увеличении обобщенной оценки результатов тестов на 1 балл для индивида с обобщенной оценкой равной 50,15 вероятность получения аттестата почти не увеличивается.

Задача 11.3.

Из 750 обратившихся за ссудой в банк 250 было в ней отказано. Для оценки вероятности получения ссуды были оценены линейная и пробит модели:

$Y = 0.5 + 1.5X$,

$$P\{Y_i = 1\} = F(Z_i),\ F(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-t^2/2} dt$$

$Z = 0.45 + 3X$

где $Y_i = 1$ для получивших ссуду и 0 иначе, X – доход просителя.

По пробит модели найти предельный эффект дохода в среднем.

Решение:

Для начала найдем средний доход просителя из линейной модели.

$\overline{Y} = 0.5 + 1.5\overline{X}$

Y_i= 1, если ссуда была получена.

Y_i= 0, иначе.

Y_i= 0 в 250 случаях => Y_i= 1 в 500 случаях.

Отсюда $\overline{Y} = \frac{500}{750} = \frac{2}{3}$.

Подставим это значение в $\overline{Y} = 0.5 + 1.5\overline{X}$:

$$\frac{2}{3} = \frac{1}{2} + 1{,}5 \cdot \overline{X} => \overline{X} = 0{,}111$$

Теперь найдем предельный эффект в пробит модели:

$$z(\overline{x}) = 0{,}45 + 3 \cdot 0{,}111 = 0{,}783333$$

$$f(z(\overline{x})) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} = 0{,}2935$$

$$\frac{\partial p(\overline{x})}{\partial x} = \frac{\partial p}{\partial z} \cdot \frac{\partial z}{\partial x} = f(z(\overline{x})) \cdot \hat{\beta}_j \approx 0{,}2935 \cdot 3 \approx 0{,}88$$

При увеличении дохода просителя на 1 единицу (в данной задаче доход измеряется, скорее всего в крупных единицах, например в млн. руб.) для индивида со средним доходом вероятность получения ссуды увеличивается на 88%.

Задача 11.4.

Для того, чтобы определить, эффективна ли новая методика преподавания микроэкономики, провели следующий эксперимент: протестировали всех студентов по микроэкономике в конце первого и второго семестра. Часть студентов во втором семестре обучали по новой методике, часть по старой. После этого в качестве объясняющей выбрали переменную Y, равную 1, если результат студента улучшился и 0 в противном случае, а в качестве объясняющих переменных X_1 – результаты

теста в первом семестре, X_2 – средний балл по остальным предметам, D – равную 1, если студент обучался по новой методике и 0, если по старой.

По имеющимся данным оценили пробит- модель:

$$P\{Y_i = 1\} = F(Z_i),\ F(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-t^2/2} dt$$

$$Z = -7.452 + 0.052X_1 + 1.626X_2 + 1.426D.$$

Найдите предельный эффект переменной D при средних значениях $\overline{X}_1 = 21.938$,

$\overline{X}_2 = 3.117$ (разность вероятностей улучшения результата при D = 1 и D = 0).

Решение:

Рассчитаем предельный эффект переменной D при средних значениях:

$p(z(\overline{x}_1, \overline{x}_2; D = 1)) - p(z(\overline{x}_1, \overline{x}_2; D = 0))$.

$$z(\overline{x}_1, \overline{x}_2; D = 1) = -7{,}452 + 0{,}052 \cdot 21{,}938 + 1{,}626 \cdot 3{,}117 + 1{,}426 \cdot 1 = 0{,}183$$

$$z(\overline{x}_1, \overline{x}_2; D = 0) = -7{,}452 + 0{,}052 \cdot 21{,}938 + 1{,}626 \cdot 3{,}117 + 1{,}426 \cdot 0 = -1{,}243$$

$$p(z(\overline{x}_1, \overline{x}_2; D = 1)) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{0{,}183} e^{-\frac{t^2}{2}} dt = \frac{1}{\sqrt{2\pi}} e^{-\frac{0{,}183^2}{2}} = 0{,}392$$

$$p(z(\overline{x}_1, \overline{x}_2; D = 0)) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{-1{,}243} e^{-\frac{t^2}{2}} dt = \frac{1}{\sqrt{2\pi}} e^{-\frac{-1{,}243^2}{2}} = 0{,}184$$

$$p(z(\overline{x}_1, \overline{x}_2; D = 1)) - p(z(\overline{x}_1, \overline{x}_2; D = 0)) = 0{,}392 - 0{,}184 \approx 0{,}2$$

Переход на обучение по новой методике увеличивает вероятность улучшения результата студента на 20%.

Задание 11.3.

Для анализа аудитории, использующей Интернет для учебы по данным для 1314 индивидов были оценены линейная и пробит модели (последняя с предельными эффектами), в которых intlear = 1 при использовании индивидом Интернета для учебы и 0 в противном случае, male = 1 для мужчин и 0 для женщин, income – заработная плата индивида по основному месту работы, age – возраст. В скобках указаны соответственно t или z - статистики. В чем состоят недостатки линейной модели? Дайте интерпретацию полученным результатам.

$$INTLEAR = \underset{(18.48)}{-0.78} - \underset{(-11.45)}{.013}\ AGE - \underset{(-0.96)}{4.53 \cdot 10^{-10}}\ INCOME - \underset{(-3.05)}{0.073}\ MALE$$

$$P\{INTLEAR_i = 1\} = F(Z_i),\ F(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-t^2/2} dt$$

$$Z = \underset{(7.25)}{1.01} - \underset{(-10.82)}{0.044}\, AGE - \underset{(-0.98)}{1.51\cdot 10^{-9}}\, INCOME - \underset{(-3.08)}{0.23}\, MALE$$

Marginal effects after probit

variable	dy/dx	Std. Err.	z	P>z	[95% C.I.]
age	-.0147444	.00133	-11.08	0.000	-.017353 -.012136
income	-5.07e-10	.00000	-0.98	0.328	-1.5e-09 5.1e-10
male*	-.0783057	.02536	-3.09	0.002	-.128015 -.028597

(*) dy/dx is for discrete change of dummy variable from 0 to 1

2) Борзых, Демешев, задача 6.12.

Методом максимального правдоподобия оценили логит-модель $\hat{y}_i^* = 2 + 3x_i - 5z_i$

1. Оцените вероятность того, что $y_i = 1$ для $\bar{x} = 5, \bar{z} = 3.5$.
2. Оцените предельный эффект увеличения x на единицу на вероятность того, что $y_i = 1$ для $\bar{x} = 5, \bar{z} = 3.5$.
3. При каком значении x предельный эффект увеличения z на 1 в точке $\bar{z} = 3.5$ будет максимальным?

4. По 1000 наблюдений Винни-Пух оценил логистическую модель $\mathbb{P}(Y_i = 1) = F(\beta_0 + \beta_1 X_i)$, где X_i — количество времени в часах, проведённое в гостях, а Y_i — факт застревания при выходе.

   Оценки параметров равны $\hat{\beta}_0 = 1$, $\hat{\beta}_1 = 2$, с оценкой ковариационной матрицы

$$\begin{pmatrix} 0.25 & 0.1 \\ 0.1 & 0.36 \end{pmatrix}.$$

   а) Проверьте значимость отдельных коэффициентов при уровне значимости 5%;

   б) Найдите предельный эффект времени, проведённого в гостях, на вероятность застрять при выходе для получасового визита;

   в) Найдите максимально возможный предельный эффект.